A Summary of Fairness in Machine Learning

Leonid Shpaner[1]

[1]University of California, Los Angeles, lshpaner@ucla.edu

Abstract

Biased evaluations and decisions stemming from algorithms have the potential to disproportionately affect various demographic groups. As artificial intelligence becomes more integrated into critical decision-making processes, it is imperative to address these biases to ensure fair and equitable outcomes. This paper discusses the nature of algorithmic bias, the impact of such biases on real-world applications, and the effectiveness of various fairness metrics in mitigating these biases. By exploring the relationships between demographic parity, equalized odds, and Pareto efficiency, the paper highlights the challenges and necessary considerations in designing unbiased AI systems. The goal is to foster a deeper understanding of how to balance fairness and utility in machine learning applications, ensuring that these technologies contribute positively to society.

Keywords: Bias and Fairness, Demographic Parity, Algorithmic Transparency, Ethical AI, Machine Learning Ethics, Fair Decision Making

1 Algorithmic Bias

Algorithmic bias refers to the presence of inequity in the predictions or decisions made by an AI system or algorithm. In the context of binary decision outcomes, algorithmic bias can appear in various forms, resulting in some groups being unfairly impacted by the system's decisions. Binary decision outcomes involve two possible results: positive (1) or negative (0). For instance, an AI system designed to accept or reject loan applications may assign a positive outcome (1) to loan approval and a negative outcome (0) to loan rejection. Several factors can introduce algorithmic bias into the system, including skewed training data, prejudiced labels, or inherent issues in the algorithm. Discriminatory training data can result from historical disparities or systemic biases that affect specific demographic groups. When the algorithm is trained on this prejudiced data, it might unintentionally perpetuate the existing biases in its decisions.

A healthcare-related example of algorithmic bias can be found in predictive algorithms used for allocating medical resources, such as identifying high-risk patients who may need additional care or interventions. These algorithms analyze patient demographics, medical history, and other factors to predict outcomes and assist healthcare providers in decision-making.

However, algorithmic bias can emerge in these predictive healthcare models if the training data or algorithms used are biased due to inadequate representation of diverse health conditions or demographic groups. For instance, if a model is trained on a dataset that over-represents a particular demographic group or under-represents certain health conditions, it may develop biases that affect its predictions and recommendations.

Regarding binary decision outcomes, algorithmic bias can result in a disparate impact on distinct demographic groups.

Disproportionate false positives occur when a specific group receives more positive outcomes than warranted, not reflective of their actual merits or qualifications. For instance, an algorithm may approve loans for individuals from a particular demographic group at a higher rate, even if they present a higher default risk.

Disproportionate false negatives can also occur when a group experiences more negative outcomes than warranted. For example, in loan applications, an algorithm might disproportionately reject eligible applicants from a specific demographic due to biased data.

To address algorithmic bias in binary decision outcomes, it is essential to consider fairness metrics such as demographic parity—a condition where decisions are independent of the protected demographic features, ensuring equal acceptance rates across different groups. By integrating these fairness principles and actively working to reduce biases, we can develop more just algorithms that yield fairer decisions for all parties involved.

Example: Classifying Kidney Transplant Eligibility

This example assesses kidney transplant eligibility based on critical factors such as blood pressure, age, height, and weight, which are crucial for determining a patient's suitability for surgery and ensuring successful post-transplant outcomes. Additionally, while sex and race may inform assessments of genetic predispositions and specific health risks, these factors are incorporated with a commitment to ethical standards and equity. The decision is primarily driven by medical evaluations and organ compatibility, adhering to guidelines that ensure practices are scientifically justified and ethically sound. The binary decision, denoted as B, indicates whether a patient is approved (Yes = 1) or not approved (No = 0) for a kidney transplant.

The expression $\mathbb{E}[D] \neq \mathbb{E}[D|A]$ illustrates that the expected decision D across the general population does not equal the expected decision within specific demographic groups A (e.g., defined by race). This disparity suggests a violation of demographic parity, indicating that decisions are not made independently of race, thereby disadvantaging or favoring certain groups. Such insights drive the need for continuous review and adaptation of eligibility criteria, ensuring they are based on the most current scientific evidence and ethical considerations.

The critical outcome to consider is patient survival following the transplant decision. These eligibility criteria are integrated within a broader framework of patient-centered care and public health goals, aiming to reduce the burden of kidney disease while improving overall patient outcomes. The multi-disciplinary transplant teams, including nephrologists, surgeons, and social workers, ensure that all aspects of the patient's health and social context are considered, promoting fairness and efficiency in the use of scarce medical resources like donor organs.

Furthermore, It is crucial to ensure that the integration of race, sex, and other sensitive attributes in assessments, which addresses genetic factors affecting disease trajectory and organ compatibility, does not lead to systematic bias. Ongoing training in cultural competence and involving patients in decision-making processes are essential to maintain transparency and uphold ethical standards in healthcare. This approach not only improves transplant outcomes but also ensures equitable healthcare practices.

2 Group Dependent Algorithmic Bias

Expanding on the concept of algorithmic bias, this section examines how disparities in false positive and false negative rates can significantly impact different demographic groups. In situations where biases disproportionately affect certain groups, these differences can lead to unjust outcomes by exacerbating inequalities within the healthcare system.

Equalized odds is a fairness metric that necessitates both true positive rates (TPR) and false positive rates (FPR) to be equal across all demographic groups. It ensures that the algorithm does not perform differently for different demographic groups, thereby upholding fairness in outcomes. The failure to meet this metric, highlighted by the inequality $\mathbb{E} \neq \mathbb{E}[R|O, A]$ indicates that expected outcomes $\mathbb{E}$ vary based on race R, outcome O, and demographic group A. This variation points to potential systemic biases that can influence algorithmic decisions, emphasizing a crucial area for ongoing scrutiny and improvement.

Biases in decision-making systems, particularly those related to life-critical applications such as kidney transplant eligibility, can have profound consequences. For instance, if certain racial groups face systematic disadvantages due to biased algorithms, this could translate into significant disparities in treatment outcomes, including higher mortality rates or longer wait times for organ transplants. Such biases challenge the ethical foundations of medical decision-making and necessitate robust, continuous revisions to ensure all applications in healthcare adhere to ethical standards and do not perpetuate or exacerbate health inequities.

One demographic group experiencing a higher TPR than another, meaning a higher percentage of true positive cases, points to an underlying issue where certain groups are favored over others. Similarly, if a higher percentage of false negative cases are identified within one group, this could lead to undue negative consequences for individuals in that group, further compounding the issue of fairness.

Conversely, One demographic group experience a higher FPR than another, meaning that a higher percentage of false positive cases are identified within that group. This could lead to undue negative consequences for individuals in that group, even if they should not have been subjected to those consequences based on their true status.

Separation, another fairness criterion, is also critical in addressing biases. It requires that the scores or probabilities generated by an algorithm be conditionally independent of protected attributes (e.g., demographic group), given the true outcomes. This ensures that the treatment of individuals within different demographic groups is fair and unbiased, even if the algorithms and true outcomes are complex. The separation criterion requires that the decision scores or probabilities output by a model be statistically independent of any protected attributes when conditioned on the actual outcome. Mathematically, this can be described using conditional probabilities:

The separation criterion can be mathematically expressed as:

$$P(\hat{Y} = y|A = a, Y = y') = P(\hat{Y} = y|Y = y') \tag{1}$$

Here, $\hat{Y}$ represents the predicted outcome by the model, A is the protected attribute, and Y is the actual outcome. This criterion requires that the probability distribution of the predictions $\hat{Y}$ should be the same across different groups defined by A for each actual outcome Y.

After discussing the theoretical basis of the separation criterion, the following algorithm demonstrates how this criterion might be implemented to ensure fair treatment across different demographic groups, as defined by protected attributes. The necessary structures are initialized and the conditional distributions needed to verify whether the separation criterion is met effectively are computed.

Algorithm 1 Check Separation Criterion

```
Initialize conditional_distributions as an empty dictionary
for each outcome y' in true_outcomes do
    Extract sub_predictions, sub_attributes where true_outcomes = y'
    for each attribute a in sub_attributes do
        Extract group_predictions where sub_attributes = a
        Compute the distribution of group_predictions
        Store this distribution in conditional_distributions[y'][a]
    end for
end for
satisfied ← True
for each outcome y' in true_outcomes do
    Compute the overall distribution of sub_predictions for y'
    for each attribute a in conditional_distributions[y'] do
        if distribution significantly differs from the overall distribution then
            satisfied ← False
            break
        end if
    end for
    if not satisfied then
        break
    end if
end for
return satisfied
```

To address these ongoing issues of bias and fairness, implementing comprehensive machine learning techniques, re-sampling methods, and adjusting decision thresholds are vital. This approach not only aims to correct for disparities but also to develop algorithms that can consistently ensure fair treatment across all demographic groups.

The primary objective remains to achieve a consistent expected outcome across all demographic groups, thereby fostering fair and accurate decision-making that respects the diversity and rights of all individuals.

3 Equality of Outcome vs. Demographic Parity

Equality of outcome is a fairness metric that aims to ensure that each demographic group achieves the same rate of positive outcomes, regardless of the underlying differences between groups. However, enforcing equality of outcome may lead to relative harm in some cases, specifically through under-acceptance.

Under-acceptance occurs when individuals who deserve a positive outcome (e.g., qualified job applicants or loan recipients) are not given the opportunity they merit due to the enforcement of equality of outcome. This can happen if a higher-performing demographic group is held back to achieve equal outcomes with a lower-performing group, causing members of the higher-performing group to be unfairly denied opportunities they would have otherwise received.

Demographic parity, on the other hand, is a fairness metric that focuses on ensuring equal acceptance rates across demographic groups, without considering individual qualifications or performance. Demographic parity ensures that each group receives the same proportion of positive outcomes, without making adjustments based on the group's qualifications or abilities. Since demographic parity does not involve redistributing outcomes based on group performance, it does not create under-acceptance in the same way as equality of outcome. However, it is important to note that demographic parity does not account for differences in qualifications between groups, which can lead to other forms of unfairness. For example, demographic parity might result in over-acceptance for a less-qualified group or under-acceptance for a more-qualified group, as it only focuses on achieving equal acceptance rates.

While demographic parity avoids under-acceptance issues associated with equality of outcome, it may not provide a complete fairness solution, as it does not consider the differences in qualifications between demographic groups. It is essential to carefully select and apply fairness metrics that are appropriate for the specific context and goals of a given AI system or algorithm.

4 Maximizing Marginal Utility Given Demographic Parity Constraint

To maximize marginal utility with a demographic parity constraint, we set up a utility function and introduce a Lagrange multiplier to incorporate the constraint.

Let's denote utility as $U(x, y)$ where x and y represent the inputs or features of two demographic groups A and B.

Demographic parity constraint states that the selection rates for each group should be the same, which can be expressed as:

$$P(\text{decision} = 1|\text{group} = A) = P(\text{decision} = 1|\text{group} = B) \tag{2}$$

Let's assume that the decision boundaries for both groups can be represented by linear equations:

$$f_{A(x)} = ax + b \tag{3}$$
$$f_{B(y)} = cy + d \tag{4}$$

Let $P_A = P(\text{decision} = 1 \mid \text{group} = A)$ and $P_B = P(\text{decision} = 1 \mid \text{group} = B)$.

Then, the Lagrangian $\mathscr{L}$ can be written as:

$$\mathscr{L}(x, y, \lambda) = U(x, y) - \lambda(P_A - P_B) \tag{5}$$

Now we need to take the partial derivatives with respect to x, y and λ and set them to zero:

$$\frac{\partial \mathscr{L}}{\partial x} = \frac{\partial U}{\partial x} - \frac{\lambda \partial(P(\text{decision} = 1|\text{group} = A))}{\partial x} = 0 \tag{6}$$

$$\frac{\partial \mathscr{L}}{\partial y} = \frac{\partial U}{\partial y} - \frac{\lambda \partial(P(\text{decision} = 1|\text{group} = A))}{\partial y} = 0 \tag{7}$$

$$\frac{\partial \mathscr{L}}{\partial \lambda} = P(\text{decision} = 1|\text{group} = A) - P(\text{decision} = 1|\text{group} = B) = 0 \tag{8}$$

These equations form a system of nonlinear equations which can be solved for x, y, and λ. Solving for these variables will provide the optimal allocation that maximizes utility under the demographic parity constraint. Keep in mind that finding an explicit solution may be challenging or impossible for more complex utility functions or decision boundaries. In such cases, numerical optimization methods, like gradient descent, can be employed.

5 Demographic Parity vs. Pareto Efficiency

Kearns et al. (2018) explore the concept of subgroup fairness in machine learning algorithms and discuss the trade-offs between fairness, utility, and Pareto efficiency in the context of fairness constraints. The authors suggest various techniques for auditing and learning subgroup fairness while taking into account multiple fairness metrics such as demographic parity, ultimately aiming to achieve a balance between fairness and efficiency (Kearns et al., 2018).

Let's consider a simple hypothetical example with two demographic groups, A and B. Suppose we have a limited number of resources (e.g., job positions) that we need to allocate to these groups.

Let x be the proportion of resources allocated to group A and $(1 - x)$ be the proportion allocated to group B. For simplicity, let's assume that the utility derived from allocating resources to each group is given by:

$$U_{A(x)} = ax \tag{9}$$

$$U_{B(x)} = b(1 - x) \tag{10}$$

where a and b are the constants representing the marginal utilities per unit of resource for groups A and B, respectively.

Under demographic parity, we want to equalize the proportion of resources allocated to both groups.

$$x = 1 - x \tag{11}$$

Solving for x, we find that $x = 0.5$. This means that we would allocate 50% of the resources to group A and the remaining 50% to group B to achieve demographic parity.

Now, let's consider Pareto efficiency. To achieve Pareto efficiency, we want to maximize the total utility without making any individual worse off. In this case, we want to maximize the sum of the utilities for both groups:

$$U_{Total(x)} = U_{A(x)} + U_{B(x)} = U_{B(x)} = ax + b(1 - x) \tag{12}$$

To find the optimal resource allocation that maximizes U_Total, we can find the critical points by taking the derivative of U_Total with respect to x and setting it equal to zero:

$$\frac{\partial U_{Total}}{\partial x} = a - b = 0 \tag{13}$$

Solving for a and b, we find that $a = b$.

However, in many real-world scenarios, it is unlikely that $a = b$, as the marginal utilities for different demographic groups may not be equal. In such cases, allocating resources based on demographic parity ($x = 0.5$) would not result in Pareto efficiency, as the total utility could be increased by reallocating resources from the group with lower marginal utility to the group with higher marginal utility.

This example illustrates that achieving demographic parity does not guarantee Pareto efficiency, a state where no individual's utility can be improved without worsening another's, emphasizing optimal resource allocation without harm to any party. In practice, the relationship between demographic parity and Pareto efficiency can be more complex due to various factors, such as different utility functions or additional constraints. Nonetheless, the example highlights the potential trade-offs between fairness and efficiency in decision-making processes.

6 Conclusion

This paper has examined the multifaceted nature of algorithmic bias and its implications in machine learning systems, focusing on the critical need for fairness metrics such as demographic parity, equalized odds, and equal opportunity. Through a series of examples and theoretical discussions, we have highlighted how biases in algorithms can disproportionately impact different demographic groups, leading to unfair decision outcomes.

The exploration into fairness metrics emphasizes their essential role in creating more equitable AI systems. However, we also recognize that achieving absolute fairness is complex and often involves trade-offs, such as those between demographic parity and Pareto efficiency. This complexity is evident in our discussions on maximizing utility under fairness constraints, where demographic parity can conflict with optimal resource allocation.

Furthermore, the paper suggests that while demographic parity aims to equalize outcomes across groups without regard to individual qualifications, it may not always result in fair or efficient outcomes. This calls for a nuanced approach to implementing fairness metrics, considering the specific contexts and goals of each AI system.

Future work in this field should continue to refine these metrics and explore new methods for mitigating bias, with an emphasis on practical applications and the development of tools for auditing and adjusting algorithms in real-world scenarios. By advancing our understanding and technology in this area, we can better ensure that AI systems perform justly across all segments of society, truly supporting the ideals of equity and fairness.

In conclusion, while significant strides have been made in understanding and addressing algorithmic bias, the journey towards truly fair AI systems is ongoing. It is a collective challenge that requires continuous effort from researchers, practitioners, and policymakers alike.

References

[1] Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., & Zhang, L. (2016). Deep Learning with Differential Privacy. *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security (CCS '16)*, 308-318. https://doi.org/10.1145/2976749.2978318

[2] Kearns, M., Neel, S., Roth, A., & Wu, Z. S. (2018). Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness. In J. Dy & A. Krause (Eds.), *Proceedings of the 35th International Conference on Machine Learning*, 80, 2564–2572

[3] Liu, Lydia [Simonds Institute] (2022, November 9). *What Really Matters for Fairness in Machine Learning: Delayed Impact and Other Desiderata* [Video]. YouTube. https://www.youtube.com/watch?v=P1SBnDTylko&t=211s